Invariant ordering of item-total regressions

Tijmstra, J.; Hessen, D.J.; van der Heijden, P.G.M.; Sijtsma, K.

Published in:
Psychometrika

DOI:
10.1007/s11336-011-9201-0

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

PSYCHOMETRIKA—VOL. 76, NO. 2, 217–227
APRIL 2011
DOI: 10.1007/S11336-011-9201-0

INVARIANT ORDERING OF ITEM-TOTAL REGRESSIONS

JESPER TIJMSTRA, DAVID J. HESSEN, AND PETER G.M. VAN DER HEIJDEN

UTRECHT UNIVERSITY

KLAAS SIJTSMA

TILBURG UNIVERSITY

A new observable consequence of the property of invariant item ordering is presented, which holds under Mokken's double monotonicity model for dichotomous data. The observable consequence is an invariant ordering of the item-total regressions. Kendall's measure of concordance W and a weighted version of this measure are proposed as measures for this property. Karabatsos and Sheu proposed a Bayesian procedure (Appl. Psychol. Meas. 28:110–125, 2004), which can be used to determine whether the property of an invariant ordering of the item-total regressions should be rejected for a set of items. An example is presented to illustrate the application of the procedures to empirical data.

Key words: double monotonicity Mokken model, invariant item ordering, invariant ordering of item-total regressions, Kendall's W, manifest invariant item ordering, nonparametric item response theory.

1. Introduction

Mokken's double monotonicity (DM) model (1971) is a nonparametric item response theory (IRT) model for the ordinal measurement of a single latent variable by means of a set of dichotomously scored items. The DM model is characterized by its definition of the item response function (IRF), which relates the probability of giving a positive or correct response to the latent variable. Unlike in parametric IRT models, such as the Rasch model (1960) and the Birnbaum models (1968), in the DM model the IRF is not parametrically defined. Instead, only order constraints are placed on the IRFs of a set of items. In contexts where the assumption of a parametric definition of the IRF is questionable or for purposes where an ordinal measurement level is sufficient, the DM model may be preferred over parametric IRT models.

Four assumptions define the DM model (Mokken, 1971; Sijtsma and Molenaar, 2002). The first assumption is unidimensionality, stating that the items measure one common latent variable. The second assumption restricts the item scores to be independent given the latent variable, and is known as the assumption of local independence (LI). The third assumption is latent monotonicity, stating that each IRF is a monotone nondecreasing function of the latent variable. The fourth assumption specifies that IRFs are nonintersecting, which is also known as invariant item ordering (IIO; Sijtsma and Junker, 1996).

The assumption of IIO distinguishes the DM model from the monotone homogeneity model (Mokken, 1971), in which IIO is not assumed, but which does assume unidimensionality, LI and latent monotonicity. Sijtsma and Junker (1996) discuss a variety of situations in which IIO is desirable or even necessary, such as the use of starting and stopping rules in intelligence testing based on the order of the item difficulties, or the analysis of differential item functioning. They point out that the property is also useful in the context of person-fit analysis, where IIO can greatly facilitate the detection of aberrant response patterns.

Requests for reprints should be sent to Jesper Tijmstra, Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands. E-mail: j.tijmstra@uu.nl

More generally, the absence of IIO may complicate the interpretation of test results. For example, if a particular item is more difficult for high-ability subjects than another item, but easier for low-ability subjects, it may be difficult to provide an explanation as to why this difference in item ordering exists. For this reason, having IIO facilitates the interpretation of test results. Thus, IIO may be desirable when designing items for a test, and could also be used as a criterion for selecting items for a test during test construction.

Some procedures have been proposed that produce an overall measure of IIO. For example, Sijtsma and Meijer (1992) proposed to evaluate IIO using scalability coefficient H^T, which is based on Loevinger's H. The authors suggest guidelines to determine whether IIO should be rejected for a test based on this measure, but the level of significance and the power of this procedure are hard to establish. Ligtvoet, Van der Ark, Te Marvelde and Sijtsma (2010) further extended this procedure. Alternatively, Scheiblechner (2003) proposed to evaluate IIO using the weak order index σ, based on Goodman and Kruskal's γ (Goodman and Kruskal, 1954). A drawback of this approach for testing for IIO is that it investigates evidence in favor of IIO (i.e., against independence of the item scores), and does not aim at testing for violations of IIO.

Rosenbaum (1987) proved that, given LI, IIO of items i and j implies a manifest IIO of those items in any subpopulation that can be specified using the remaining items. Rosenbaum proposed to group subjects based on their unweighted sumscore on the remainder of the items. However, other groupings have also been proposed. Mokken (1971) proposed to use the score on a single item to group subjects. Another option is to use a vector containing a subset of the item scores on the remaining items, and test for IIO by investigating whether increasingness in transposition holds, which is implied by IIO (see Rosenbaum, 1987; Sijtsma and Junker, 1996). If this property holds, observing a response vector with a score of 1 on an easy item and a score of 0 on a more difficult item should always be at least as probable as observing a response vector in which the more difficult item receives a 1 and the easier item a 0.

The drawback of methods that use subgroupings based on one or more of the remaining items is that the conclusion whether IIO holds for a set of items depends on many partial results. This renders such procedures laborious and, more importantly, makes it difficult to conclude whether IIO holds for the test as a whole. For example, if one tests for IIO per item pair by making use of the restscore for the item pair under consideration (i.e., the unweighted sumscore on the remainder of the items), one has to evaluate all item pairs separately in order to evaluate IIO for the whole test, because the partitioning of subjects based on their restscores is likely to differ for different item pairs.

The use of the unweighted *total* score instead of the restscore would remedy this problem of having too many partial results, since a partitioning based on the total score would not vary over different item pairs. Thus, using the total score to test for IIO would facilitate the evaluation of the ordering of all items simultaneously instead of having to deal with separate item pairs, and would provide a more efficient method for IIO assessment. Using the total score to test for IIO would amount to determining whether the item probabilities conditional on the total score – the so-called item-total regressions – have the same ordering at every level of the total score. Such an ordering constitutes a manifest version of IIO, that is, a manifest invariant item ordering (MIIO) over the total score. Note that although an MIIO could be investigated over a variety of manifest scores, such as the mentioned restscore, in the remainder of the article the term MIIO will solely be used to refer to an MIIO over the total score.

An approach that focuses on the total score was proposed by Karabatsos and Sheu (2004), who suggested a Bayesian procedure that can be used to determine whether it is likely that an invariant ordering of the item-total regressions holds. The procedure makes use of a Gibbs sampler and results in a posterior-predictive p-value, which indicates whether we should reject or retain the assumption that the item probabilities are invariantly ordered over the total score. However, although they present this procedure as providing a test for Mokken's DM model, they

do not provide a proof that MIIO holds under the DM Mokken model. While it has already been established that an invariant ordering of the item-*rest* regressions holds under the DM Mokken model (Sijtsma and Junker, 1996), this has not yet been proven for the item-*total* regressions.

The current article presents a proof showing that the DM model for dichotomous data implies an invariant ordering of the item-total regressions – an MIIO over the total score. Therefore, statistical tests for an invariant ordering of the item-total regressions can be used to test for the DM model. This proof provides the basis for the method to test the DM model that was proposed by Karabatsos and Sheu (2004). In addition to the proof, two measures of the property of an invariant ordering of the item-total regressions are proposed, both of which make use of Kendall's (1939) measure of concordance W. The method of Karabatsos and Sheu is discussed as well, since with that method a decision can be made to reject or retain MIIO. The application of the measures based on Kendall's W and the Karabatsos and Sheu method is illustrated using two sets of empirical data.

2. Theorem and Proof

Let X_i denote the random variable for the score on item i, with realization x_i. Let $\mathbf{X} = (X_1, \ldots, X_k)$ denote the vector of item-score variables on a test with k dichotomous items, with realization $\mathbf{x} = (x_1, \ldots, x_k)$. The latent variable is denoted by θ, and the IRF of item i is denoted by $P(X_i = 1|\theta) = P_i(\theta)$. One of the assumptions of the DM model is LI of the item scores given θ, which can be presented as

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{i=1}^{k} P(X_i = x_i|\theta). \tag{1}$$

The assumption of IIO states that the items can be ordered such that

$$P_1(\theta) \leq \cdots \leq P_k(\theta), \quad \text{for all } \theta, \tag{2}$$

where the indices $1, \ldots, k$ are assigned to the items based on decreasing difficulty. Furthermore, the total score $\sum_{i=1}^{k} X_i$ is denoted by T, and its realization by t. It is the claim of the present paper that (1) and (2) together imply the observable consequence of MIIO. This is stated in the following theorem.

Theorem. *LI in* (1) *and IIO in* (2) *together imply*

$$P(X_1 = 1|T = t) \leq \cdots \leq P(X_k = 1|T = t), \quad \textit{for all } t, \tag{3}$$

where $P(X_i = 1|T = t)$ *is the item-total regression of item* i*, that is, the probability of a positive response to item* i *given* $T = t$.

Proof: Following Hessen (2005), the probability of $\mathbf{X} = \mathbf{x}$ can be written as

$$P(\mathbf{X} = \mathbf{x}|\theta) = \left\{\prod_{i=1}^{k} Q_i(\theta)\right\}\left\{\prod_{i=1}^{k-1} \omega_{ik}(\theta)^{x_i}\right\} V_k(\theta)^t,$$

where $Q_i(\theta) = 1 - P_i(\theta)$, $\omega_{ik}(\theta) = \frac{P_i(\theta)Q_k(\theta)}{Q_i(\theta)P_k(\theta)}$, and $V_k(\theta) = \frac{P_k(\theta)}{Q_k(\theta)}$. Any of the k items can serve as the reference item. So, here, the choice of item k as the reference item is arbitrary. Let A_t be

the set of all item-score vectors with total score t; that is, $A_t = \{\mathbf{x} : \sum_{i=1}^{k} x_i = t\}$. Then, we may write

$$P(T = t|\theta) = \left\{\prod_{i=1}^{k} Q_i(\theta)\right\}\left\{\sum_{A_t}\prod_{i=1}^{k-1}\omega_{ik}(\theta)^{x_i}\right\}V_k(\theta)^t,$$

and

$$P(\mathbf{X} = \mathbf{x}|T = t, \theta) = \frac{\prod_{i=1}^{k-1}\omega_{ik}(\theta)^{x_i}}{\sum_{A_t}\prod_{i=1}^{k-1}\omega_{ik}(\theta)^{x_i}}. \tag{4}$$

Also, let $B_{tx_jx_k}$ be the subset of A_t in which the scores on items j and k are x_j and x_k, respectively, and where j is an arbitrarily selected item. That is, $B_{tx_jx_k} = \{\mathbf{x} : x_j, x_k, \sum_{i=1}^{k} x_i = t\}$. Using (4), it follows that

$$P(X_j = x_j, X_k = x_k|T = t, \theta) = \frac{\omega_{jk}(\theta)^{x_j}\sum_{B_{tx_jx_k}}\prod_{i\neq j}^{k-1}\omega_{ik}(\theta)^{x_i}}{\sum_{A_t}\prod_{i=1}^{k-1}\omega_{ik}(\theta)^{x_i}},$$

for $i = 1, \ldots, k-1$, all t and θ. Note that $\sum_{B_{tx_jx_k}}\prod_{i\neq j}^{k-1}\omega_{ik}(\theta)^{x_i}$ is the same for $\{\mathbf{x} : x_j = 1, x_k = 0, \sum_{i=1}^{k} x_i = t\}$ and $\{\mathbf{x} : x_j = 0, x_k = 1, \sum_{i=1}^{k} x_i = t\}$. Hence,

$$\frac{P(X_j = 1, X_k = 0|T = t, \theta)}{P(X_j = 0, X_k = 1|T = t, \theta)} = \omega_{jk}(\theta),$$

for $j = 1, \ldots, k-1$, all t and θ. Since items j and k were selected arbitrarily, it follows that

$$P(X_i = 1, X_j = 0|T = t, \theta) = P(X_i = 0, X_j = 1|T = t, \theta)\,\omega_{ij}(\theta), \quad \text{for all } t \text{ and } \theta,$$

for any two items i and j. Now, IIO implies that for any item pair i and j, either $\omega_{ij}(\theta) \geq 1$ or $\omega_{ij}(\theta) \leq 1$, for all θ. So if we arbitrarily let $\omega_{ij}(\theta) \geq 1$ for all θ, then

$$P(X_i = 1, X_j = 0|T = t, \theta) \geq P(X_i = 0, X_j = 1|T = t, \theta), \quad \text{for all } t \text{ and } \theta.$$

Adding $P(X_i = 1, X_j = 1|t, \theta)$ to both sides gives

$$P(X_i = 1|T = t, \theta) \geq P(X_j = 1|T = t, \theta), \quad \text{for all } t \text{ and } \theta.$$

Averaging both sides with respect to an arbitrary conditional density function $f(\theta|T = t)$ yields

$$\int P(X_i = 1|T = t, \theta)\,f(\theta|T = t)\,d\theta \geq \int P(X_j = 1|T = t, \theta)\,f(\theta|T = t)\,d\theta, \quad \text{for all } t,$$

resulting in

$$P(X_i = 1|T = t) \geq P(X_j = 1|T = t), \quad \text{for all } t.$$

Hence, if IIO holds, then Equation (3) holds. This completes the proof. □

3. Evaluating an Invariant Ordering of the Item-Total Regressions

Equation (3) specifies an order restriction on the item probabilities for each value of the total score. If for each value of the total score we investigate the ordering of the proportions

TABLE 1.
Observed proportions of positive responses.

	t		
Item	1	$\cdots$	$k-1$
1	p_{11}	$\cdots$	$p_{1(k-1)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
k	p_{k1}	$\cdots$	$p_{k(k-1)}$

TABLE 2.
Proportions of positive responses at different levels of the total score, based on fictional data.

	t			
Item	1	2	3	4
1	0.10	0.19	0.38	0.63
2	0.08	0.44	0.46	0.69
3	0.20	0.21	0.58	0.81
4	0.14	0.31	0.65	0.88
5	0.48	0.85	0.92	1.00

of positive responses that were observed in the data, information can be obtained as to whether or not MIIO is violated. Table 1 displays these proportions of positive responses, denoted p_{it}, which are obtained by dividing the number of positive responses on item i at level t (denoted s_{it}) by the total number of responses at level t (denoted n_t). Note that since for $t = 0$ and $t = k$ by definition all items have the same proportion of success, the ordering at those levels does not contain information regarding MIIO and hence need not be considered. The extent to which the orderings of p_{it} $(i = 1, \ldots, k)$ differ for different values of t $(t = 1, \ldots, k-1)$ gives a rough picture of whether the data indicate that MIIO is violated. In the next subsection, it is shown how this information can be summarized using two simple measures that are based on Kendall and Babington Smith's (1939) measure of concordance W. Since the theoretical null distribution of these measures under MIIO is unknown, actually testing for MIIO using these measures is not an option that is available. In the second subsection, the statistical testing procedure proposed by Karabatsos and Sheu is discussed. This procedure results in a decision to retain or reject MIIO, and hence, whether the data contain evidence that IIO should be rejected.

3.1. Kendall's W

A measure of the correspondence between the orderings of the item probabilities at different levels of the total score provides insight into whether MIIO is violated. For this purpose, we propose the use of Kendall's measure of concordance (Kendall's W; Kendall and Babington Smith, 1939). This nonparametric measure is related to the Spearman rank correlation coefficient (Spearman, 1904), but unlike the latter, W can compare more than two orderings simultaneously. It can therefore be used to evaluate the degree of correspondence or concordance between the orderings of the items at different levels of the total score.

Kendall's W takes on values between 0 (no correspondence) to 1 (perfect ordinal correspondence), and is usually employed to compare the orderings of the ratings of different judges. Kendall's W is calculated on the basis of the rankings provided by a number of raters who independently ordered a number of objects. In the context of evaluating MIIO, these raters are replaced by the different levels of the total score (excluding $t = 0$ and $t = k$), and the objects by the items. For example, if the fictional proportions displayed in Table 2 are taken as the result of

TABLE 3.
Order of the proportions of positive responses at different levels of the total score, corresponding to Table 2.

	t				
Item	1	2	3	4	SR_i
1	2	1	1	1	5
2	1	4	2	2	9
3	4	2	3	3	12
4	3	3	4	4	14
5	5	5	5	5	20

the ratings, they can be ordered for each level of the total score, which results in the rankings in Table 3.

To determine the value of Kendall's W, the sum of the rankings of each item needs to be determined. Let R_{it} be the rank of item i obtained for total score t, where this ranking is based on increasing proportions. The sum of these rankings for item i is obtained through

$$SR_i = \sum_{t=1}^{k-1} R_{it}. \tag{5}$$

Let $\mathbf{SR} = (SR_1, \ldots, SR_k)$. We consider the deviations of the elements of **SR** from their average value. Let S represent the sum of squares of these deviations; then W can be calculated using

$$W = \frac{12S}{(k-1)^2(k^3-k)}. \tag{6}$$

For the example in Table 3 one obtains

$$W = \frac{12 \times 126}{16(125-5)} = 0.7875.$$

One possible drawback of using Kendall's W to measure the extent to which the data are in accordance with MIIO is that it does not take the number of observations into account that are available at the different levels of the total score. Kendall's W expresses the degree of correspondence between the $k-1$ item orderings, and each level receives equal weight. However, when relatively few observations are available at a specific level of the total score, observing a discordant item ordering due to chance is more likely.

The different levels of the total score can be weighted by taking into account the number of persons that have the same total score t, denoted by n_t. This can be accomplished by reformulating Equation (5) as

$$SR_i = \frac{k-1}{\sum_{t=1}^{k-1} n_t} \sum_{t=1}^{k-1} n_t R_{it}.$$

That is, the rankings that constitute SR_i are weighted by n_t, and the value of W under this weighting approach can be calculated as described in Equation (6). This weighted version of Kendall's W is expected to be less sensitive to random fluctuations in the orderings.

If the sample value of W or its weighted counterpart equals 1, there is no evidence available that MIIO is violated. If the value is smaller than 1, one can conclude that the ordering of the proportion of successes is not the same at every level of the total score. However, since deviations from perfect correspondence could be due to chance, one cannot determine whether MIIO is

violated based on W alone. To determine whether MIIO should be rejected, information is needed about the likelihood of obtaining a value as extreme (i.e., as low) as the observed value, given that MIIO holds.

Unfortunately, the only null distribution of W or the weighted version of W that is available is the distribution that corresponds to independence of the item orderings at the different levels of the total score. This null distribution enables one to test whether the item orderings show more invariance than one would expect to find under independence. However, if one wants to test for *violations* of MIIO, another null distribution is needed, one that corresponds to the situation where the item orderings are invariant over the total score. Regrettably, W does not have a theoretical null distribution corresponding to the assumption of MIIO, and hence there is no exact test available that can be used to determine the probability of obtaining a value as low as the observed value. Thus, the two measures provide a useful way of summarizing the extent to which the data are in accordance with MIIO, but they do not provide an exact test to decide whether the property of MIIO – and hence the DM model – should be rejected.

3.2. Karabatsos and Sheu's Posterior-Predictive p-Value

To test for MIIO, Karabatsos and Sheu (2004) proposed using the table with observed proportions (Table 1), and determining how likely these proportions are when one assumes that MIIO holds. Let $\mathbf{p}$ denote the $k \times (k-1)$ matrix containing the observed proportions of positive responses as in Table 1; that is

$$\mathbf{p} = (p_{it} | i = 1, \ldots, k; t = 1, \ldots, k-1),$$

where p_{it} denotes the observed proportion corresponding to item i and a total score t. Let $\mathbf{\Delta}$ denote the $k \times (k-1)$ matrix containing the item-total regressions; that is

$$\mathbf{\Delta} = (P_{it} | i = 1, \ldots, k; t = 1, \ldots, k-1),$$

where $P_{it} = P(X_i = 1 | T = t)$, and hence $\mathbf{\Delta} \in (0, 1)^{k(k-1)}$. The likelihood of the data $\mathbf{p}$ given $\mathbf{\Delta}$ can then be assumed to be a product of $k(k-1)$ independent binomial probability mass functions:

$$L(\mathbf{p}|\mathbf{\Delta}) = \prod_{i=1}^{k} \prod_{t=1}^{k-1} \binom{n_t}{s_{it}} P_{it}^{s_{it}} (1 - P_{it})^{n_t - s_{it}}. \tag{7}$$

Let $\pi(\mathbf{\Delta})$ denote the prior distribution of the item probabilities in $\mathbf{\Delta}$. Let $\mathbf{\Omega}$ denote the subset of $(0, 1)^{k(k-1)}$ that is in accordance with Equation (3); that is, the set of all matrices $\mathbf{\Delta}$ for which MIIO holds. The order constraints that follow from MIIO restrict this prior distribution in the following way:

$$\pi(\mathbf{\Delta}) \begin{cases} > 0 & \text{iff } \mathbf{\Delta} \in \mathbf{\Omega}, \\ = 0 & \text{iff } \mathbf{\Delta} \notin \mathbf{\Omega}. \end{cases} \tag{8}$$

By combining Equations (7) and (8), the order-constrained posterior distribution of $\mathbf{\Delta}$ can be obtained through

$$\pi(\mathbf{\Delta}|\mathbf{p}) = \frac{L(\mathbf{p}|\mathbf{\Delta})\pi(\mathbf{\Delta})}{\int_{\mathbf{\Omega}} L(\mathbf{p}|\mathbf{\Delta})\pi(\mathbf{\Delta}) d\mathbf{\Delta}}. \tag{9}$$

Equation (9) cannot be evaluated analytically, but one can make use of the unconstrained posterior distribution of $\mathbf{\Delta}$,

$$\pi(\mathbf{\Delta}_u|\mathbf{p}) = \frac{L(\mathbf{p}|\mathbf{\Delta}_u)\pi(\mathbf{\Delta}_u)}{\int L(\mathbf{p}|\mathbf{\Delta}_u)\pi(\mathbf{\Delta}_u) d\mathbf{\Delta}_u},$$

where the subscript u indicates that $\boldsymbol{\Delta}$ is no longer constrained to be part of $\boldsymbol{\Omega}$. This unconstrained posterior distribution can be modeled in terms of a beta distribution. Here, one has to use a beta density prior $\pi(\boldsymbol{\Delta}_u)$ in which each probability P_{it} is specified independently and without the restriction of MIIO.

Using a Gibbs sampler (for details, see Karabatsos and Sheu, 2004), a large number of samples $(\boldsymbol{\Delta}^r | r = 1, \ldots, R)$ can be generated from the order-constrained posterior distribution $\pi(\boldsymbol{\Delta}|\mathbf{p})$. After discarding a proper burn-in period, $1, \ldots, B$, these draws result in an approximation of the posterior distribution $\pi(\boldsymbol{\Delta}|\mathbf{p})$.

To test whether MIIO is violated, the observed proportions $\mathbf{p}$ can be compared to the posterior-predictive distribution,

$$\pi\left(\mathbf{p}^{\text{rep}}|\mathbf{p}\right) = \int_{\boldsymbol{\Omega}} \pi\left(\mathbf{p}^{\text{rep}}|\boldsymbol{\Delta}\right)\pi(\boldsymbol{\Delta}|\mathbf{p})\, d\boldsymbol{\Delta}.$$

This latter distribution can be approximated using the same Gibbs sampler, since after each iteration r, $\boldsymbol{\Delta}^{(r)}$ can be used to generate a new set of data, $\mathbf{p}^{\text{rep}}$. The posterior-predictive distribution is then approximated by the set of $\mathbf{p}^{\text{rep}}$ obtained in the Gibbs sampler.

To use this posterior-predictive distribution to test for violations of MIIO, Karabatsos and Sheu (2004) proposed using a chi-square discrepancy measure, defined as

$$\chi^2(\mathbf{p};\boldsymbol{\Delta}) = \sum_{i=1}^{k}\sum_{t=1}^{k-1}\left[\frac{(n_t p_{it} - n_t P_{it})^2}{n_t P_{it}}\right].$$

Using this measure, it is possible to obtain the posterior-predictive p-value (ppp-value):

$$\begin{aligned} p(\mathbf{p}|\boldsymbol{\Delta}) &= Pr\left[\chi^2\left(\mathbf{p}^{\text{rep}};\boldsymbol{\Delta}\right) \geq \chi^2(\mathbf{p};\boldsymbol{\Delta})|\mathbf{p}\right] \\ &= \int\int_{\boldsymbol{\Omega}} I\left[\chi^2\left(\mathbf{p}^{\text{rep}};\boldsymbol{\Delta}\right) \geq \chi^2(\mathbf{p}|\boldsymbol{\Delta})\right] p\left(\mathbf{p}^{\text{rep}}|\boldsymbol{\Delta}\right) p(\boldsymbol{\Delta}|\mathbf{p})\, d\mathbf{p}^{\text{rep}}\, d\boldsymbol{\Delta}, \end{aligned}$$

where I is an indicator function that equals 1 when the data $\mathbf{p}$ show less discrepancy relative to $\boldsymbol{\Delta}$ than $\mathbf{p}^{\text{rep}}$. This ppp-value indicates how likely it is to observe data as extreme as $\mathbf{p}$, under the assumption that MIIO holds. It can approximated using the samples $(\boldsymbol{\Delta}^r | r = B+1, \ldots, R)$, resulting in

$$\frac{1}{R-B}\sum_{r=B+1}^{R} I\left\{\chi^2\left(\mathbf{p};\boldsymbol{\Delta}^{(r)}\right) \geq \chi^2\left(\mathbf{p}^{\text{rep}};\boldsymbol{\Delta}^{(r)}\right)\right\}.$$

This way, by selecting large enough values for both B and R, an approximation of the ppp-value can be obtained, which indicates whether MIIO should be rejected. In order to decide whether MIIO should be rejected, a critical value needs to be selected for this ppp-value, for which Karabatsos and Sheu suggest the value of 0.15. Thus, by applying the Gibbs sampler it is possible to test for MIIO using a Bayesian approach.

4. Application to Empirical Data

The two measures based on Kendall's W and the Karabatsos and Sheu procedure were used to evaluate MIIO for two sets of empirical data from a study in developmental psychology. One scale measuring nonaggressive behavior (seven items) and another scale measuring aggressive antisocial behavior (five items) in male adolescents (Dekovic, 2003) were analyzed. Both scales

TABLE 4.
Overall proportions of positive responses and proportions conditional on t for the nonaggressive antisocial behavior scale.

Item description	Proportion Positive	t					
		1	2	3	4	5	6
1: Disregarding parent's prohibitions	0.69	0.59	0.78	0.92	0.91	0.96	1.00
2: Missing a curfew	0.44	0.15	0.34	0.67	0.84	0.89	1.00
3: Skipping school	0.16	0.02	0.09	0.18	0.25	0.50	0.59
4: Cheating on a test	0.43	0.14	0.42	0.55	0.79	0.96	0.93
5: Fare dodging	0.26	0.05	0.18	0.27	0.58	0.57	0.93
6: Shoplifting	0.24	0.03	0.15	0.25	0.44	0.75	0.86
7: Stealing from someone	0.14	0.03	0.04	0.16	0.19	0.36	0.69

TABLE 5.
Overall proportions of positive responses and proportions conditional on t for the aggressive antisocial behavior scale.

Item description	Proportion Positive	t			
		1	2	3	4
1: Fire setting	0.10	0.10	0.19	0.38	0.63
2: Carrying a weapon	0.27	0.48	0.85	0.92	1.00
3: Threatening with a weapon	0.12	0.08	0.44	0.46	0.69
4: Beating someone	0.14	0.14	0.31	0.65	0.88
5: Street fighting	0.13	0.20	0.21	0.58	0.81

consisted of polytomous items, but a dichotomization was easy to obtain, since each item measured the occurrence of specific types of antisocial behavior during the past year: to dichotomize the items, subjects received a score of 1 if the behavior had occurred, and a score of 0 otherwise. The sample size was 504, but due to missing values eight subjects were excluded from the analysis of the nonaggressive antisocial behavior scale, and six subjects were excluded from the analysis of the aggressive antisocial behavior scale.

Tables 4 and 5 provide a description of the items, with the corresponding overall sample proportions of positive responses and the sample proportions of positive responses for each level of the total score. The tables show that scores of 1 were obtained more often on the items of the nonaggressive antisocial behavior scale than on the items of the aggressive antisocial behavior scale, and that the overall proportion of positive responses varied more between the items on the former scale than on the latter one, indicating a larger spread in item difficulties on the nonaggressive behavior scale.

For both scales, the two measures based on W were calculated, and the Bayesian procedure was used to test whether MIIO should be rejected. For the nonaggressive antisocial behavior scale, we found $W = 0.923$, and the weighted version of W equaled 0.938. These values are close to 1, suggesting little evidence that MIIO is violated. The ppp-value of 0.537 (based on 5,000 iterations) supports this conclusion, and hence MIIO was not rejected for this scale.

For the scale measuring aggressive antisocial behavior, $W = 0.788$, and the weighted W equaled 0.731. These values are lower than the values obtained for the nonaggressive scale. The Bayesian procedure resulted in a ppp-value of 0.142, which is just below the critical value of 0.15. Thus, the results suggest that for this scale MIIO appears to be violated. Hence, IIO and the DM model can be rejected for this scale.

Thus, for the nonaggressive antisocial behavior scale, MIIO was not rejected, and hence IIO need not be rejected for this scale. If IIO does indeed hold for this scale, this means that the different types of behavior measured by the nonaggressive scale come in a specific order, with

one kind of behavior always being more likely than another, regardless of how antisocial the adolescent is. This could be an interesting substantive finding, resulting from the non-rejection of IIO. Likewise, it would be interesting to know why IIO does not appear to hold for the aggressive antisocial types of behavior. Perhaps some types of behavior display nonmonotonicities, only being exhibited frequently by mildly antisocial youths. Again, such a conclusion based on IIO research could result in relevant substantive considerations.

5. Conclusion and Discussion

For a test consisting of dichotomous items, it was shown that under LI the property of IIO over the latent variable implies the property of manifest invariant ordering of the item-total regressions; that is, MIIO over the total score. This result implies that MIIO not only holds for the Rasch model (Hessen, 2005), but also for the DM model (Mokken, 1971). Thus, investigating MIIO is not only useful in the context of parametric IRT, but also in the context of nonparametric IRT. Inspection of MIIO is relatively simple, and may be an attractive method in IIO research. This way, the theorem provided in this article helps to facilitate IIO research.

The two measures of MIIO, both based on Kendall and Babington Smith's (1939) measure of concordance W, reflect the extent to which the data are in accordance with MIIO. Values of 1 imply that there is no evidence available that MIIO is violated. Values lower than 1 show that the data are not completely in accordance with MIIO. Karabatsos and Sheu (2004) proposed a Bayesian procedure to determine whether it is likely that MIIO is violated. This Bayesian procedure results in a decision to reject or retain MIIO for a test consisting of dichotomous items. In addition to this Bayesian approach, it would be interesting to investigate whether it is also possible to provide a frequentist test, perhaps making use of the constrained statistical inference framework (see, e.g., Silvapulle and Sen, 2005).

Regardless of which testing procedure is used, a rejection of MIIO implies a rejection of IIO, and hence testing procedures for MIIO can be used to determine whether the application of IRT models that assume IIO, such as the DM model and by implication the Rasch model, would be appropriate. Furthermore, IIO is an attractive property in itself that one could pursue during test construction, since it allows for an unambiguous ordering of the items based on their difficulty, which makes interpretation of the test results easier. Additionally, IIO may be required in applications where starting and stopping rules need to be applied, or where the items need to be presented in order of difficulty (Sijtsma and Junker, 1996). Differential item functioning analysis and person-fit analysis may also benefit from having IIO.

Knowing whether IIO holds can also be of substantive importance. Whenever IIO is violated for two specific items, groupings of respondents can be made for which the order of the probabilities of these items is reversed. This will at the very least require an explanation, telling us why it is the case that, for example, a certain item is relatively easy for high-ability examinees, perhaps pointing to some new insight that helps them deal with an item that would otherwise be relatively difficult. As shown in the example, it might also point to the possibility that some behavior ceases to become more frequent after a certain level of the latent variable has been reached, perhaps even indicating nonmonotonicities. For these reasons, determining whether MIIO and hence IIO should be rejected can be of considerable importance.

References

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.) *Statistical theories of mental test scores* (pp. 397–479). Reading: Addison-Wesley.

Dekovic, M. (2003). Aggressive and nonaggressive antisocial behavior in adolescence. *Psychological Reports*, *93*, 610–616.

Goodman, L.A., & Kruskal, W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 732–764.
Hessen, D.J. (2005). Constant latent odds-ratios models and the Mantel-Haenszel null hypothesis. *Psychometrika*, *70*, 497–516.
Karabatsos, G., & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement*, *28*, 110–125.
Kendall, M.G., & Babington Smith, B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics*, *10*, 275–287.
Ligtvoet, R., Van der Ark, L.A., Te Marvelde, J.M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*, 578–595.
Mokken, R.J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen and Lydiche.
Rosenbaum, P.R. (1987). Comparing item characteristic curves. *Psychometrika*, *52*, 217–233.
Scheiblechner, H. (2003). Nonparametric IRT: testing the bi-isotonicity of isotonic probabilistic models (ISOP). *Psychometrika*, *68*, 79–96.
Sijtsma, K., & Junker, B.W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*, 79–105.
Sijtsma, K., & Meijer, R.R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, *16*, 149–157.
Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks: SAGE Publications.
Silvapulle, M.J., & Sen, P.K. (2005). *Constrained statistical inference: inequality, order, and shape restrictions*. Hoboken: John Wiley & Sons, Inc.
Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Manuscript Received: 18 JAN 2010
Final Version Received: 20 SEP 2010
Published Online Date: 29 JAN 2011